Reinforcement Learning and Reward

Emma Brunskill
CS234
Week 10
Winter 2023

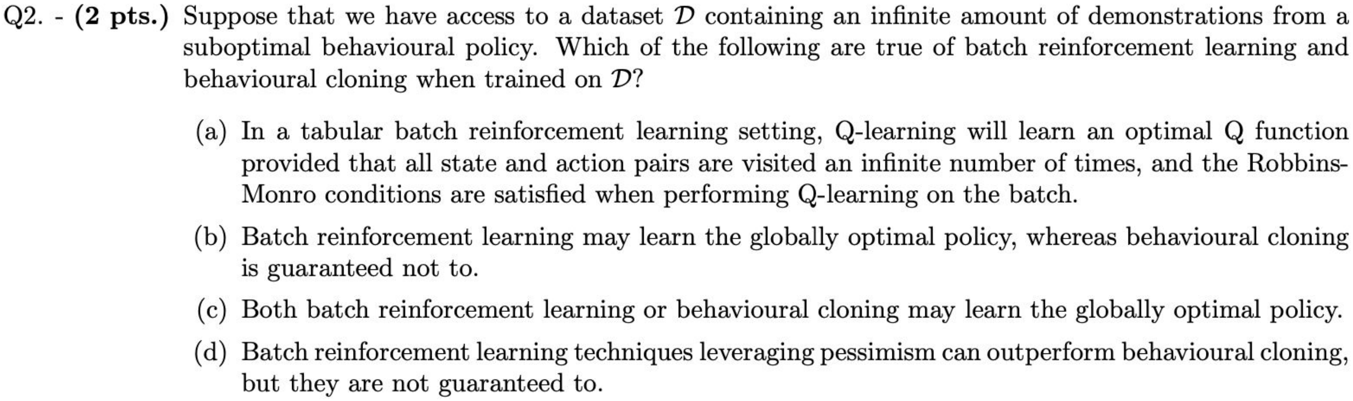
Where We Are

- Last: Learning from historical data and Quiz
- Now: Reinforcement Learning in the Wild, Course wrap up
    - Rewards, alignment
- Next: Project Presentations

Plan for today

- Quiz review
- Reward in RL

Quiz



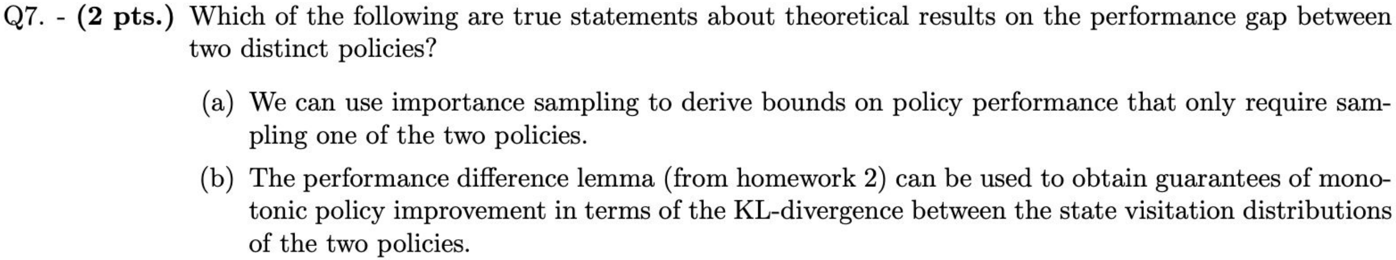
Q2. - **(2 pts.)** Suppose that we have access to a dataset $\mathcal{D}$ containing an infinite amount of demonstrations from a suboptimal behavioural policy. Which of the following are true of batch reinforcement learning and behavioural cloning when trained on $\mathcal{D}$?

(a) In a tabular batch reinforcement learning setting, Q-learning will learn an optimal Q function provided that all state and action pairs are visited an infinite number of times, and the Robbins-Monro conditions are satisfied when performing Q-learning on the batch.

(b) Batch reinforcement learning may learn the globally optimal policy, whereas behavioural cloning is guaranteed not to.

(c) Both batch reinforcement learning or behavioural cloning may learn the globally optimal policy.

(d) Batch reinforcement learning techniques leveraging pessimism can outperform behavioural cloning, but they are not guaranteed to.



Q7. - **(2 pts.)** Which of the following are true statements about theoretical results on the performance gap between two distinct policies?

(a) We can use importance sampling to derive bounds on policy performance that only require sampling one of the two policies.

(b) The performance difference lemma (from homework 2) can be used to obtain guarantees of monotonic policy improvement in terms of the KL-divergence between the state visitation distributions of the two policies.

Q8. - **(2 pts.)** For a fixed state $s \in \mathcal{S}$, which of the following are true of importance sampling for a behavior policy $\mu : \mathcal{S} \to \Delta(\mathcal{A})$, an evaluation policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, and a reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$?

(a) If $\mu(a \mid s) = 0 \implies \pi(a \mid s) = 0$ for all $a \in \mathcal{A}$, then $\mathbb{E}_{a \sim \pi(\cdot|s)}\left[\mathcal{R}(s,a)\right] = \mathbb{E}_{a \sim \mu(\cdot|s)}\left[\frac{\mu(a|s)}{\pi(a|s)}\mathcal{R}(s,a)\right]$

(b) If $\pi(a \mid s) = 0 \implies \mu(a \mid s) = 0$ for all $a \in \mathcal{A}$, then $\mathbb{E}_{a \sim \pi(\cdot|s)}\left[\mathcal{R}(s,a)\right] = \mathbb{E}_{a \sim \mu(\cdot|s)}\left[\frac{\mu(a|s)}{\pi(a|s)}\mathcal{R}(s,a)\right]$

(c) If $\mu(a \mid s) = 0 \implies \pi(a \mid s) = 0$ for all $a \in \mathcal{A}$, then $\mathbb{E}_{a \sim \pi(\cdot|s)}\left[\mathcal{R}(s,a)\right] = \mathbb{E}_{a \sim \mu(\cdot|s)}\left[\frac{\pi(a|s)}{\mu(a|s)}\mathcal{R}(s,a)\right]$

(d) Estimating the value of π using μ can yield a higher variance estimate than estimating a policy's value using on-policy rollouts and using these to compute a Monte Carlo estimate

(e) If μ explores the state-action space visited by π, importance sampling will result in a high-bias, low-variance estimator.

(f) If μ explores the state-action space visited by π, importance sampling will result in an unbiased estimator.

Q10. - **(2 pts.)** Theoretical properties. Select all that are true:

(a) Under GLIE and the Robins-Munroe conditions, tabular Q-learning is guaranteed to be a PAC RL algorithm

(b) A greedy policy in a multi-armed bandit will always suffer linear regret.

(c) UCB has a sublinear regret in a multi-armed bandit unless the reward gaps (the difference between the reward of the optimal action, and the rewards of other actions) are large.

(d) Algorithms that optimize to be probably approximately correct and algorithms that minimize regret should yield the same policy

(e) Algorithms that optimize the expected sum of rewards and algorithms that minimize regret should yield the same policy

(f) A PAC algorithm for a multi-armed bandit could have a linear regret

(g) Any PAC algorithm is also guaranteed to converge to the optimal policy

Q11. - **(2 pts.)** Consider a multi-armed bandit setting in which there are 3 possible actions and the agent observes the following rewards after taking the given actions. Rewards are binary. Assume all algorithms first pull each arm once. If an algorithm has ties, assume they can be broken randomly. Which of the following algorithms may have generated this sequence?

t (Timestep)	a_t (Action)	r_t (Observed Reward)
1	1	1
2	2	1
3	3	0
4	2	1
5	2	1

(a) ϵ-greedy, if $\epsilon = 0$

(b) ϵ-greedy, if $0 < \epsilon < 1$

(c) ϵ-greedy, if $\epsilon = 1$

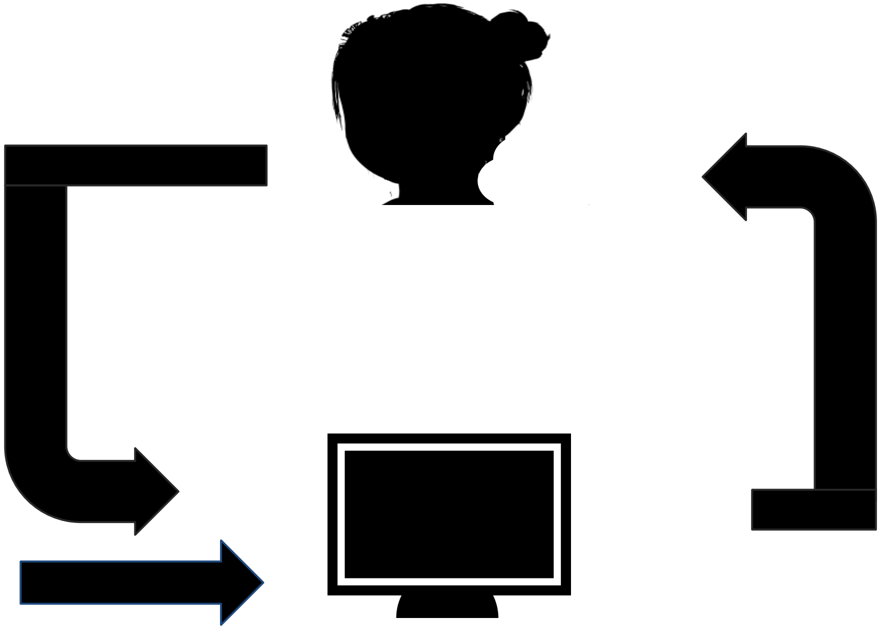
(d) Upper Confidence Bound with Hoeffding confidence intervals.

(e) Thompson Sampling

Plan for today

- Quiz review
- **Reward in RL**

Blood Pressure Management



State / Observation:

Blood pressure
Gender
Location

Action / Decision

Suggest exercise or meditation

Reward:
If in healthy range: +1
If use medication: -0.05
-

Beyond Expected Reward

- In this class focused on expected scalar reward
- In many real settings
  - Distribution of outcomes (distributional RL, conditional value at risk, …)
  - Multiple-objective (high reward and low cost and …)
  - Constrained maximization (safety, fairness, …)

nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

At last — a computer program that can beat a champion Go player PAGE 484

ALL SYSTEMS GO

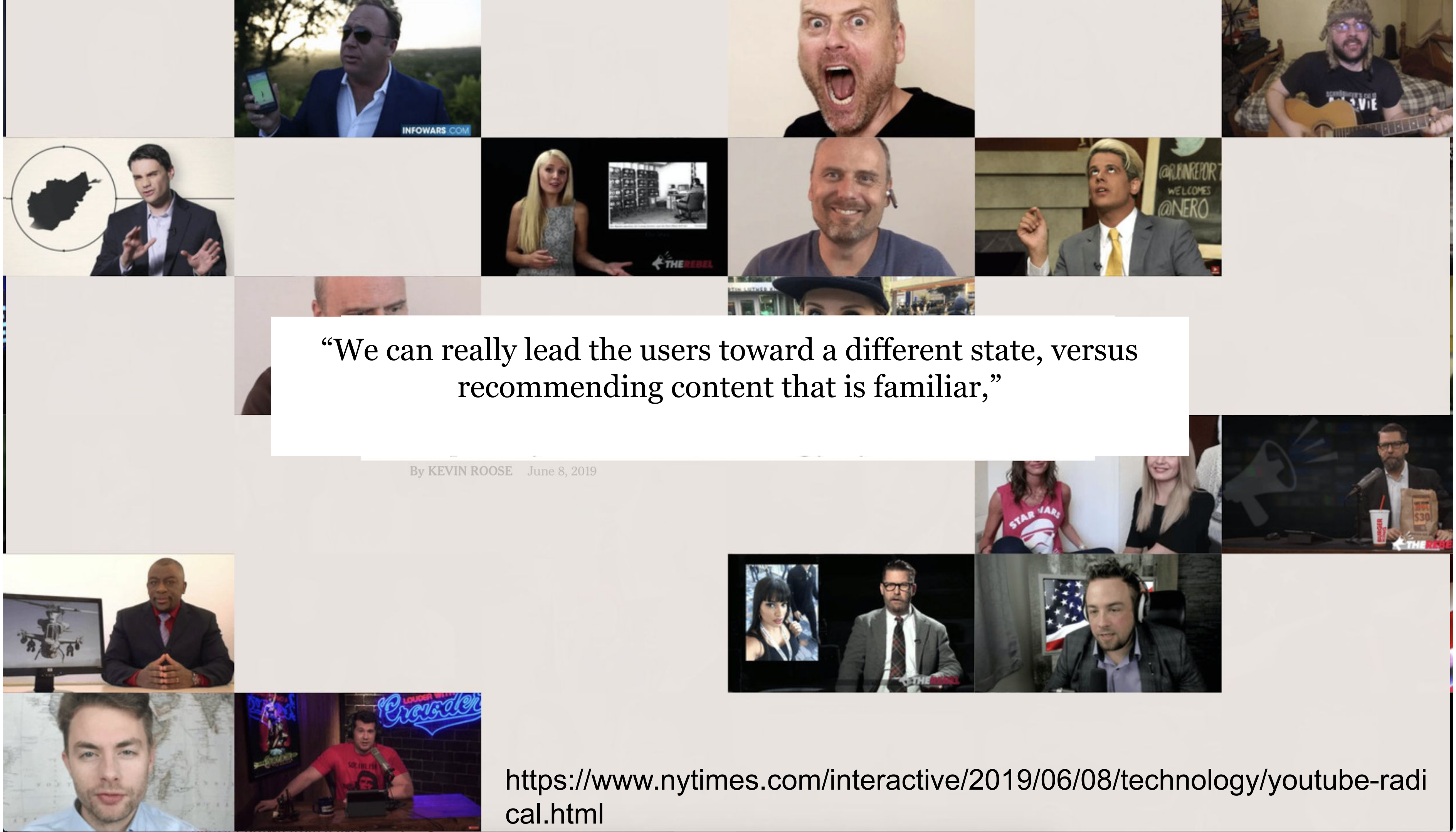
Recall Example During My 1st Lecture: AI Teacher

- Student initially does not know addition (easier) nor subtraction (harder)
- Teaching agent can provide activities about addition or subtraction
- Agent gets rewarded for student performance:
  - +1 if student gets problem right,
  - -1 if get problem wrong
- (Think/Discuss) What type of policy would a RL agent learn? Is this what the human designer of this system would likely want?

Caleb Cain was a college dropout looking for direction. He turned to YouTube.

- In last 2 years have been trying out using reinforcement learning
- "… designed to maximize users' engagement over time by predicting which recommendations would expand their tastes and get them to watch not just one more video but many more."

https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html

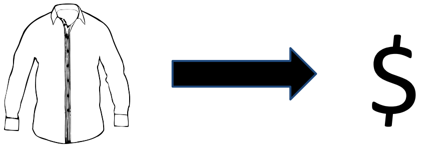


By KEVIN ROOSE June 8, 2019

> “We can really lead the users toward a different state, versus recommending content that is familiar,”

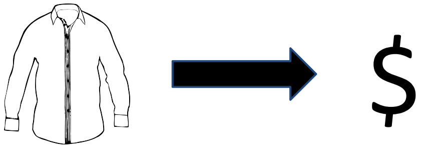
https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html

Supervised Learning



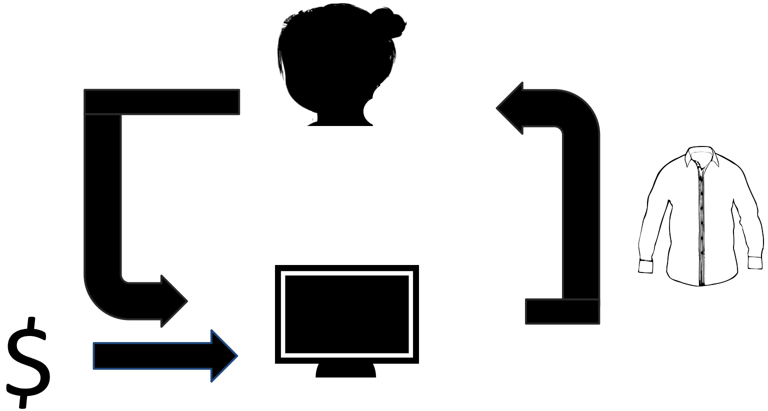
Recommend things people already like*

Supervised Learning



Recommend things people already like*

Reinforcement Learning

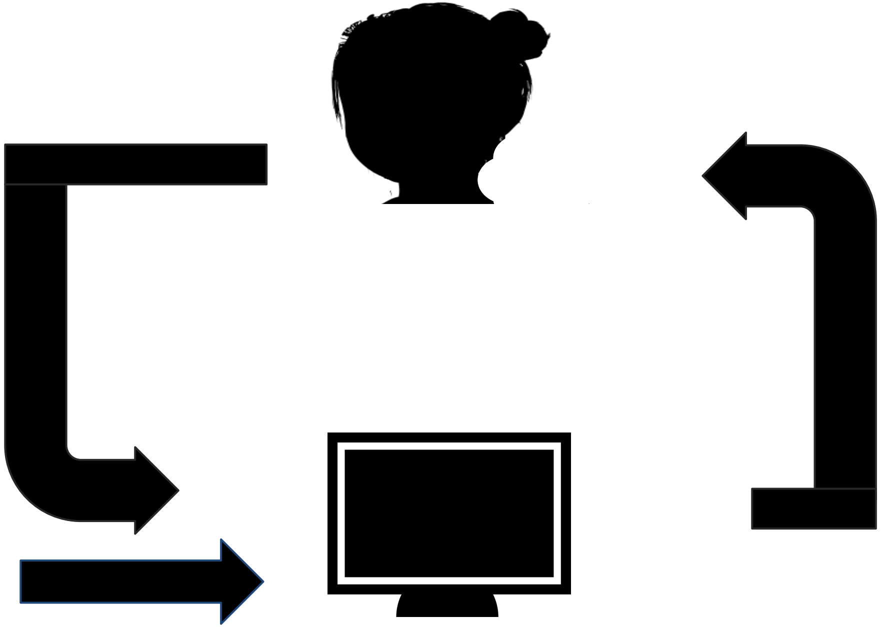


Provide recommendations so people will *(potentially change into people who)* buy more

Reinforcement Learning is Trying to Change (the State of) the World

State / Observation:

Blood pressure
Gender
Location



Action / Decision

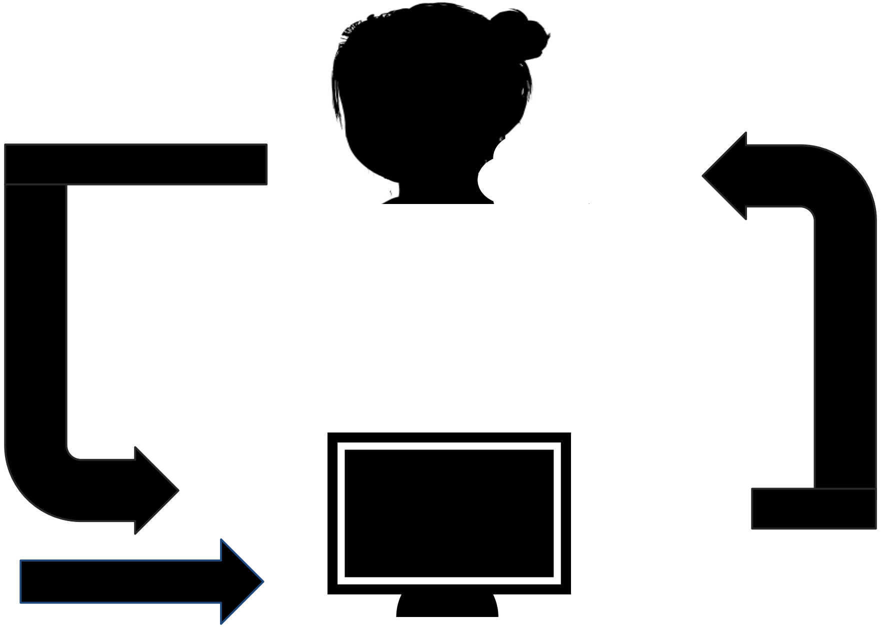
Suggest exercise or meditation

Reward:
If in healthy range: +1
If use medication: -0.05

Reinforcement Learning is Trying to Change (the State of) the World

State / Observation:

Blood pressure
Gender
Location

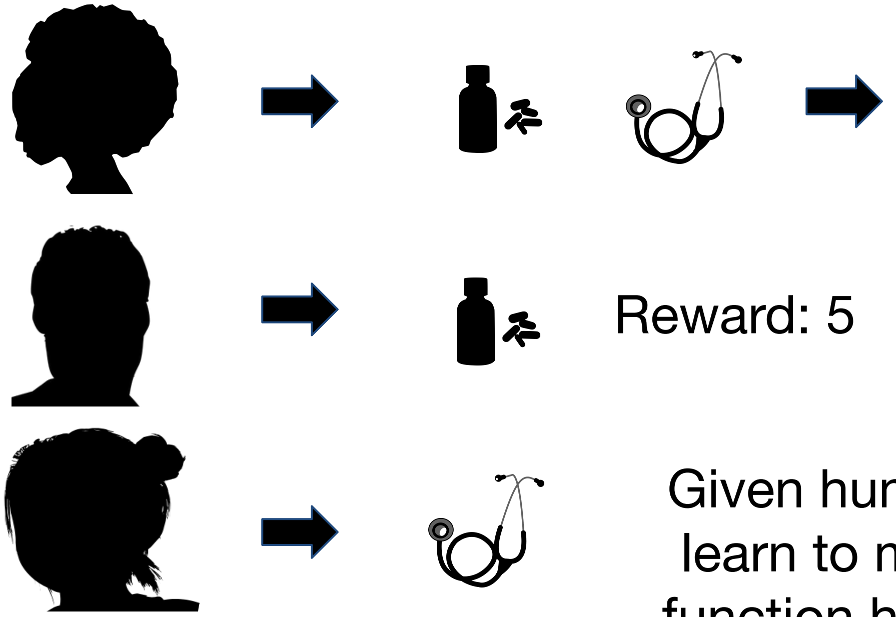


Action / Decision

Suggest exercise or meditation

What is the Reward?

One Idea: Learn the Rewards of People



Reinforcement Learning

Reward: 92

Multi-armed Bandits

Reward: 5

Imitation Learning & Inverse RL

Given human expert decisions, learn to mimic or learn reward function humans are optimizing

Value Alignment

- How can we ensure RL agent is optimizing for our desired rewards?
- Stuart Russell (recent general audience book on this broad topic is Human Compatible: AI and the Problem of Control)
- Anca Dragan, Smitha Milli, Dylan Hadfield-Menell, and others

Wrapping Up CS234

Wrapping Up CS234: Final Parts

- Normal office hours this week
- Project presentations Thursday!
  - Posters (and videos for SCPD students) should be uploaded by 5pm on gradescope
  - No late days
  - See forum post for logistic details
- Final project report due March 22 at 6pm California time (no late days)

Wrapping Up CS234: Learning Objectives

- Define the key features of reinforcement learning that distinguishes it from AI and non-interactive machine learning (as assessed by the exam).
- Given an application problem (e.g. from computer vision, robotics, etc), decide if it should be formulated as a RL problem; if yes be able to define it formally (in terms of the state space, action space, dynamics and reward model), state what algorithm (from class) is best suited for addressing it and justify your answer (as assessed by the exam).
- Implement in code common RL algorithms (as assessed by the assignments).
- Describe (list and define) multiple criteria for analyzing RL algorithms and evaluate algorithms on these metrics: e.g. regret, sample complexity, computational complexity, empirical performance, convergence, etc (as assessed by assignments and the exam).
- Describe the exploration vs exploitation challenge and compare and contrast at least two approaches for addressing this challenge (in terms of performance, scalability, complexity of implementation, and theoretical guarantees) (as assessed by an assignment and the exam).

Wrapping Up CS234: Classes to Learn More

- CS332 Advanced Survey of Reinforcement Learning (me)
- CS Deep Reinforcement Learning (Chelsea Finn)
- AA 203: Optimal and Learning-based Control (Marco Pavone)
- MS&E / EE: Ben Van Roy sometimes offers a theory-oriented RL class
- To learn more about theory of RL, Sham Kakade (UW) & Wen Sun (Cornell) taught a class based on a joint manuscript https://rltheorybook.github.io https://wensun.github.io/CS6789.html

Wrapping Up CS234: The End!

- Thank you so for your questions and participation in this class!
- Reinforcement learning is a huge field with growing impact and you’ve now mastered some of the key ideas!
- Please fill in the course evaluations-- I greatly value your feedback and knowing what helped you learn and what could use improvement helps me improve the class for later students
- Keep in touch! I love hearing about what students do next, and if there are ideas in RL they keep using or advancing
- We look forward to your project presentations!